Capstone Project
Exploratory Data Analysis
on
Airbnb Dataset

Name: Harshal Marathe

Content

Airbnb Dataset

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.
This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Dataset Look

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	name	host_id	host_name	neighbourhood_grou	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listi	availability_365
2	2539	Clean & quiet apt ho	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
3	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
4	3647	THE VILLAGE OF H/	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150	3	0			1	365
5	3831	Cozy Entire Floor of I	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
6	5022	Entire Apt: Spacious	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.1	1	0
7	5099	Large Cozy 1 BR Apa	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire home/apt	200	3	74	2019-06-22	0.59	1	129
8	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60	45	49	2017-10-05	0.4	1	0
9	5178	Large Furnished Roo	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79	2	430	2019-06-24	3.47	1	220
10	5203	Cozy Clean Guest Rc	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79	2	118	2017-07-21	0.99	1	0
11	5238	Cute & Cozy Lower E	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/apt	150	1	160	2019-06-09	1.33	4	188
12	5295	Beautiful 1br on Upp	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Entire home/apt	135	5	53	2019-06-22	0.43	1	6
13	5441	Central Manhattan/n	7989	Kate	Manhattan	Hell's Kitchen	40.76076	-73.98867	Private room	85	2	188	2019-06-23	1.5	1	39
14	5803	Lovely Room 1, Garc	9744	Laurie	Brooklyn	South Slope	40.66829	-73.98779	Private room	89	4	167	2019-06-24	1.34	3	314
15	6021	Wonderful Guest Bec	11528	Claudio	Manhattan	Upper West Side	40.79826	-73.96113	Private room	85	2	113	2019-07-05	0.91	1	333
16	6090	West Village Nest - S	11975	Alina	Manhattan	West Village	40.7353	-74.00525	Entire home/apt	120	90	27	2018-10-31	0.22	1	0
17	6848	Only 2 stops to Manh	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837	-73.95352	Entire home/apt	140	2	148	2019-06-29	1.2	1	46
18	7097	Perfect for Your Pare	17571	Jane	Brooklyn	Fort Greene	40.69169	-73.97185	Entire home/apt	215	2	198	2019-06-28	1.72	1	321
19	7322	Chelsea Perfect	18946	Doti	Manhattan	Chelsea	40.74192	-73.99501	Private room	140	1	260	2019-07-01	2.12	1	12
20	7726	Hip Historic Brownstc	20950	Adam And Charity	Brooklyn	Crown Heights	40.67592	-73.94694	Entire home/apt	99	3	53	2019-06-22	4.44	1	21
21	7750	Huge 2 BR Upper Ea	17985	Sing	Manhattan	East Harlem	40.79685	-73.94872	Entire home/apt	190	7	0			2	249
22	7801	Sweet and Spacious	21207	Chaya	Brooklyn	Williamsburg	40.71842	-73.95718	Entire home/apt	299	3	9	2011-12-28	0.07	1	0
23	8024	CBG CtyBGd HelpsH	22486	Lisel	Brooklyn	Park Slope	40.68069	-73.97706	Private room	130	2	130	2019-07-01	1.09	6	347
24	8025	CBG Helps Haiti Roo	22486	Lisel	Brooklyn	Park Slope	40.67989	-73.97798	Private room	80	1	39	2019-01-01	0.37	6	364
25	8110	CBG Helps Haiti Rm	22486	Lisel	Brooklyn	Park Slope	40.68001	-73.97865	Private room	110	2	71	2019-07-02	0.61	6	304
26	8490	MAISON DES SIREN	25183	Nathalie	Brooklyn	Bedford-Stuyvesant	40.68371	-73.94028	Entire home/apt	120	2	88	2019-06-19	0.73	2	233
27	8505	Sunny Bedroom Acro	25326	Gregory	Brooklyn	Windsor Terrace	40.65599	-73.97519	Private room	60	1	19	2019-06-23	1.37	2	85
28	8700	Magnifique Suite au	26394	Claude & Sophie	Manhattan	Inwood	40.86754	-73.92639	Private room	80	4	0			1	0
29	9357	Midtown Pied-a-terre	30193	Tommi	Manhattan	Hell's Kitchen	40.76715	-73.98533	Entire home/apt	150	10	58	2017-08-13	0.49	1	75
30	9518	SPACIOUS, LOVELY	31374	Shon	Manhattan	Inwood	40.86482	-73.92106	Private room	44	3	108	2019-06-15	1.11	3	311
31	9657	Modern 1 BR / NYC /	21904	Dana	Manhattan	East Village	40.7292	-73.98542	Entire home/apt	180	14	29	2019-04-19	0.24	1	67
32	9668	front room/double be	32294	Ssameer Or Trip	Manhattan	Harlem	40.82245	-73.95104	Private room	50	3	242	2019-06-01	2.04	3	355
33	9704	Spacious 1 bedroom	32045	Teri	Manhattan	Harlem	40.81305	-73.95466	Private room	52	2	88	2019-06-14	1.42	1	255
34	9782	Loft in Williamsburg	32169	Andrea	Brooklyn	Greenpoint	40.72219	-73.93762	Private room	55	4	197	2019-06-15	1.65	3	284
35	9783	back room/bunk beds	32294	Ssameer Or Trip	Manhattan	Harlem	40.8213	-73.95318	Private room	50	3	273	2019-07-01	2.37	3	359
36	10452	Large B&B Style roo	35935	Angela	Brooklyn	Bedford-Stuyvesant	40.6831	-73.95473	Private room	70	1	74	2019-05-12	0.66	2	269
37	10962	Lovely room 2 & gar	9744	Laurie	Brooklyn	South Slope	40.66869	-73.9878	Private room	89	4	168	2019-06-21	1.41	3	340
38	11452	Clean and Quiet in E	7355	Vt	Brooklyn	Bedford-Stuyvesant	40.68876	-73.94312	Private room	35	60	0			1	365

Dataset Summary

- ID -- ID is Dataset's Unique Identifier, which has been store as a integer datatype in our Airbnb Dataset.
- Name – In the Name Column there is room title or room name or it can be hotel name preset as a object(string) datatype.
- Host ID – In the host id column there is unique id or number present which belongs to each host.
- Host name – basically in the host name column all the host names present as the string datatype.
- Neighbourhood Group – In the neighbourhood group Column all the Group name of neighbourhood prenent as a string datatype.
- Neighbourhood -- In the neighbourhood column all the neighbourhood name present as a string datatype.
- Latitude -- Latitude is the measurement of distance north or south of the Equator. And the latitude is present as float data type.
- Longitude -- Longitude is the measurement east or west of the prime meridian. And the longitude is also present as a float data type.

- Room Type -- In the room type column different room types present like private room, shared room as a string datatype.
- Price -- price column consist the price of the neighbourhoods or rooms as a integer datatype.
- Minimum nights -- how many nights guest or host stay in the room that information store in minimun nights column as a integer datatype.
- Number of reviews -- Till the today how many reviews that host or room get that information store in the number of reviews column as a integer datatype.
- Last review -- In the last review column the recent review is store which has been given recently.
- Reviews per month -- In the review per month column how many reviews got within a month that information store month wise as float datatype.
- Calculated host listings count -- in this column listings count is present as a integer datatype.
- availability 365 -- In the availability 365 column whether the rooms is available or not that kind of information store as a integer datatype.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. with the help of statistical summary and graphical representations.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them.

Required packages

- ✓ **Pandas**
- ✓ **Numpy**
- ✓ **Matplotlib**
- ✓ **Seaborn**

Reading Dataset

```
[ ] #Mount the google drive with google colab for importing the data.
    from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Reading Dataset

```
[ ] #importing the dataset
    df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Capstone Project EDA/Airbnb NYC 2019.csv')
```

```
#First Five row of dataset
df.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_mont
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.2
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.3
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	Na
3	3831	Cozy Entire Floor of	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire	89	1	270	2019-07-05	4.6

Pandas Functions

read_csv(): read csv pandas function help us to load our dataset into notebook.

head(): head function shows us top 5 records of our dataset

tail(): unlike head() function tail function shows last 5 record of dataset

info(): info function gives us all column name with datatype information.

describe(): describe function gives us statistical summary of dataset.

isna(): isna function gives us null value information like which column having how many null values.

nunique(): nunique function gives us the values which are non unique.

Dataset Overview

```
#First Five row of dataset
df.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_mont
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.2
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.3
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	Na
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.6
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.1

Statistical Summary

```
#statistic summary of dataset
df.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Data Cleaning

Data Cleaning

```
#Removing null values from the dataset
df.dropna(inplace = True)
```

```
[ ] # Removing ID column becouse it has no any prediction power for predict dependent variable.
    df.drop(['id'], axis = 1, inplace = True)
```

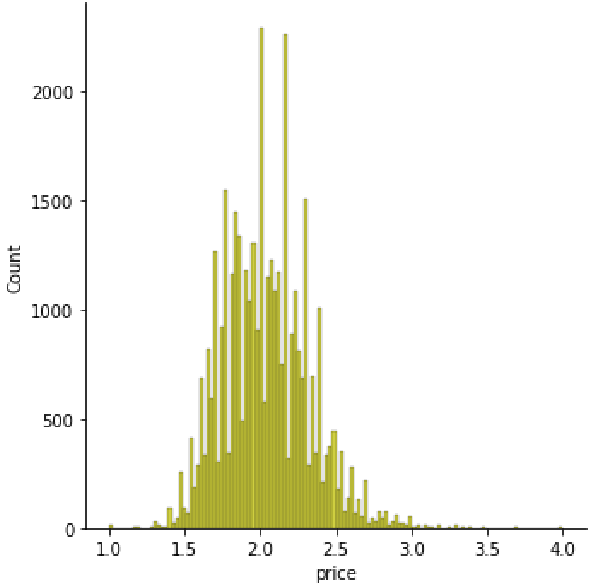
```
[ ] df.isna().sum()
```

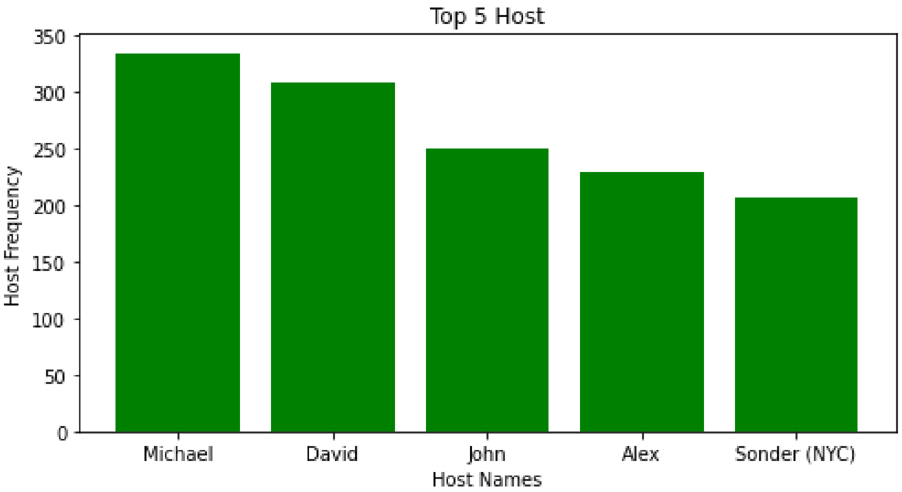
```
name                              0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

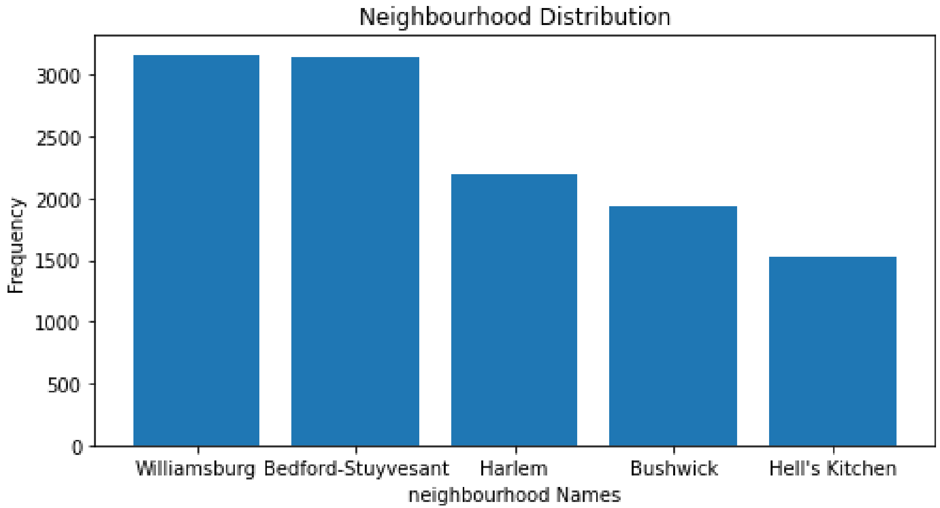
Univariate Analysis

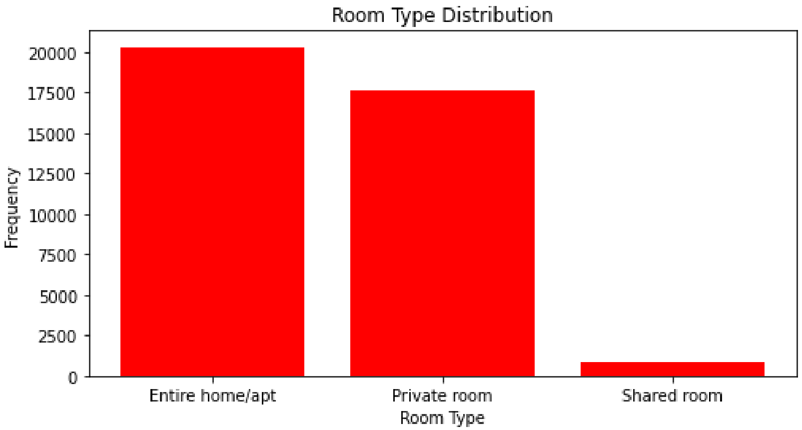
Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable.

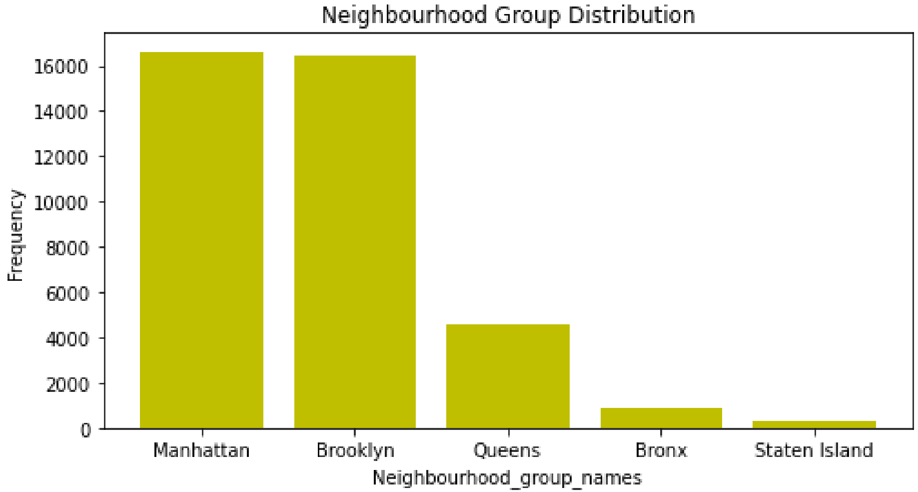
Univariate analysis is a basic kind of analysis technique for statistical data. Here the data contains just one variable and does not have to deal with the relationship of a cause and effect.





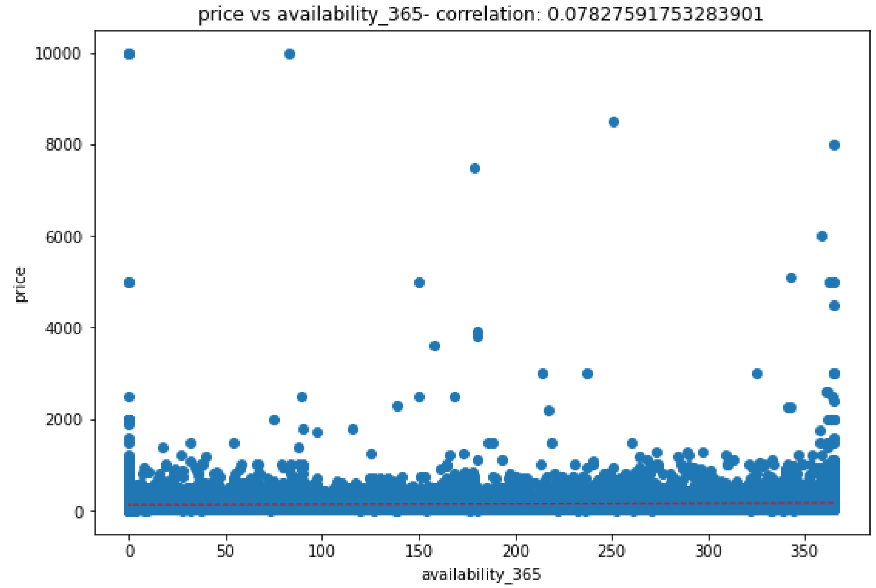

Neighbourhood Distribution
Frequency
3000
2500
2000
1500
1000
500
0
Williamsburg
Bedford-Stuyvesant
Harlem
Bushwick
Hell's Kitchen
neighbourhood Names

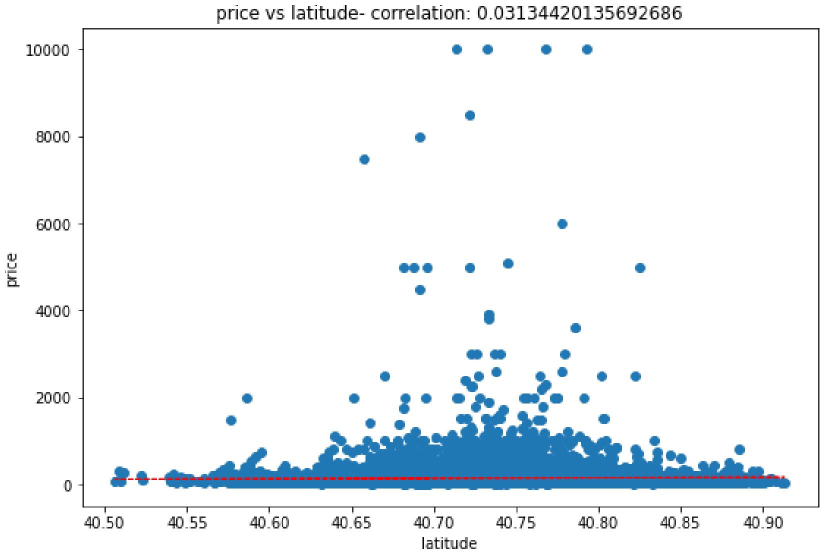

Room Type Distribution
Frequency
20000
17500
15000
12500
10000
7500
5000
2500
0
Entire home/apt
Private room
Shared room
Room Type

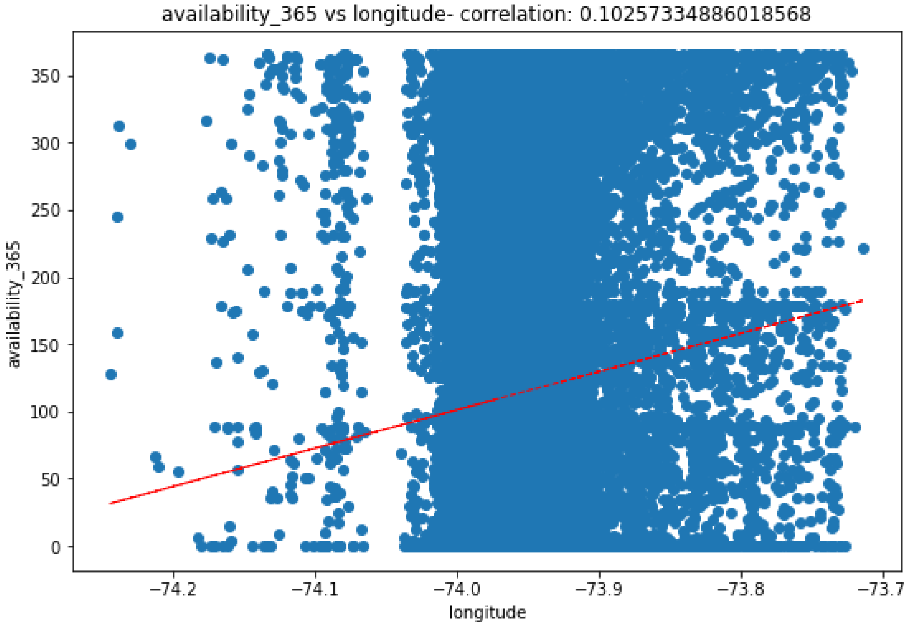

Neighbourhood Group Distribution
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
Manhattan
Brooklyn
Queens
Bronx
Staten Island
Neighbourhood_group_names

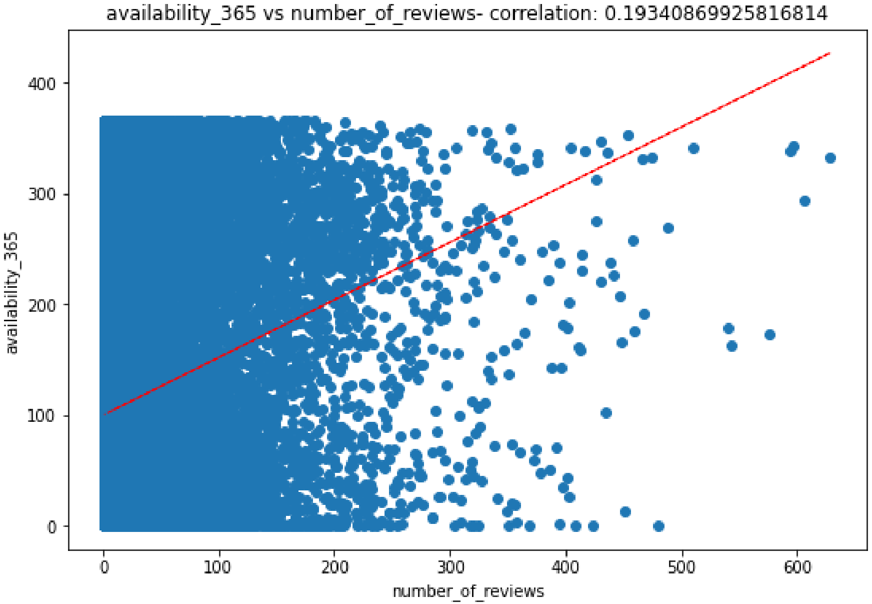
Bivariate Analysis

Bivariate analysis is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y.



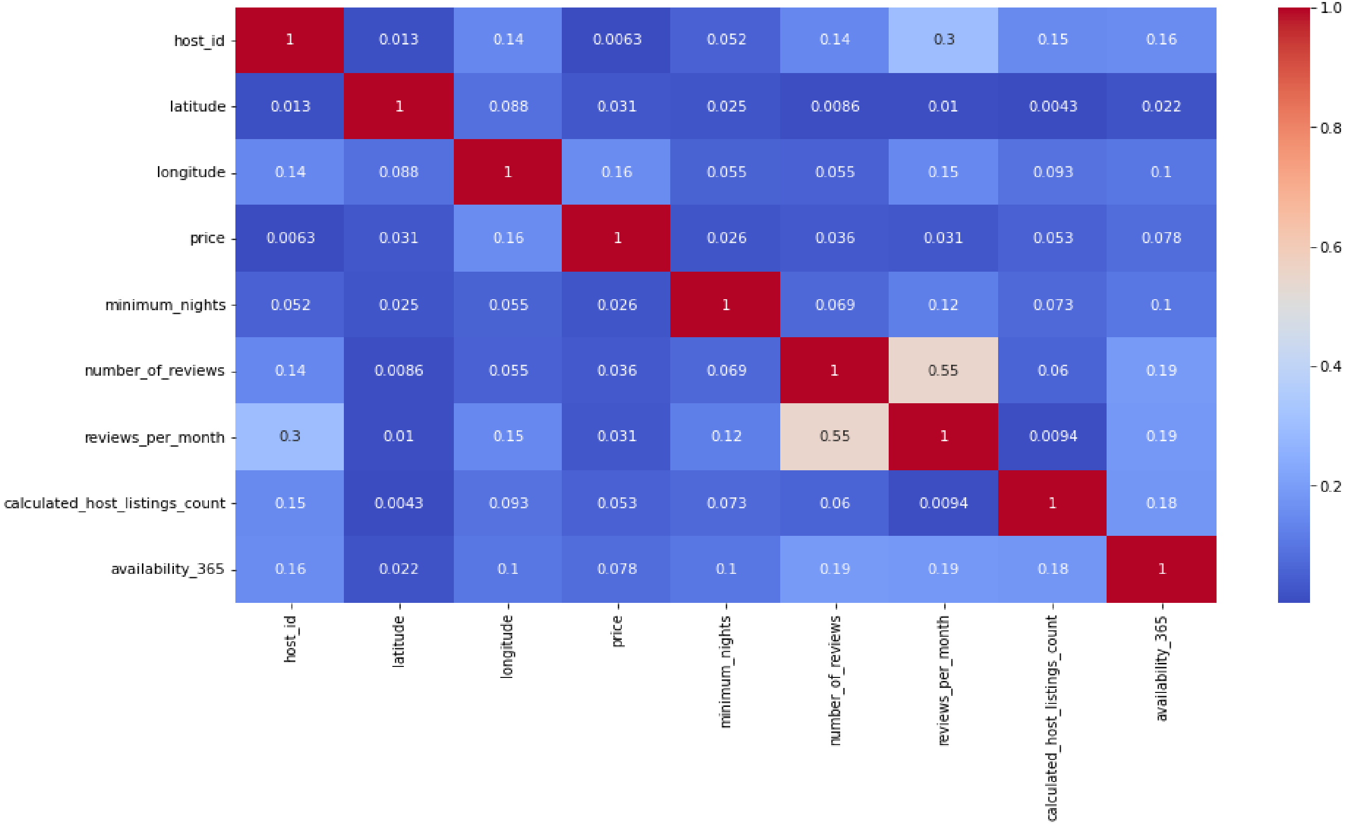



availability_365 vs longitude- correlation: 0.10257334886018568
availability_365
350
300
250
200
150
100
50
0
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
longitude

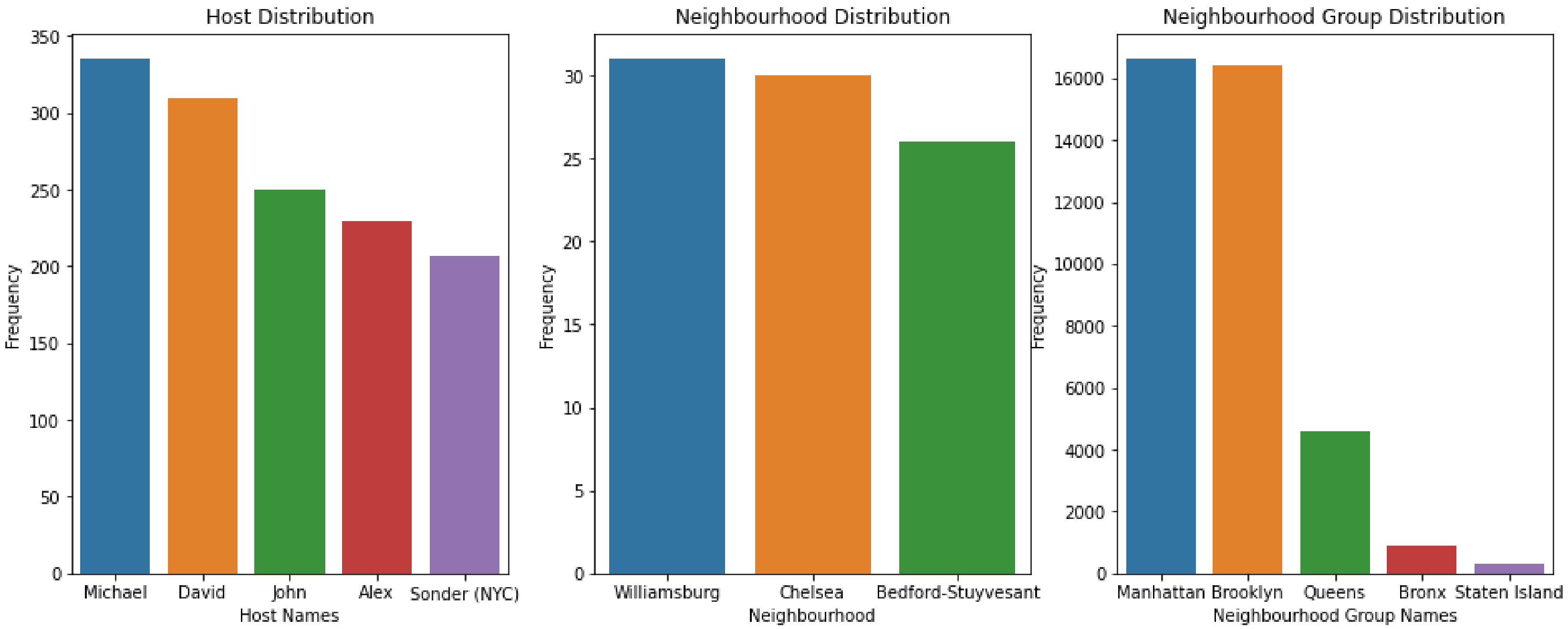

availability_365 vs number_of_reviews- correlation: 0.19340869925816814
availability_365
400
300
200
100
0
0
100
200
300
400
500
600
number_of_reviews

Multivariate Analysis

Multivariate analysis of variance (MANOVA) is used to measure the effect of multiple independent variables on two or more dependent variables.

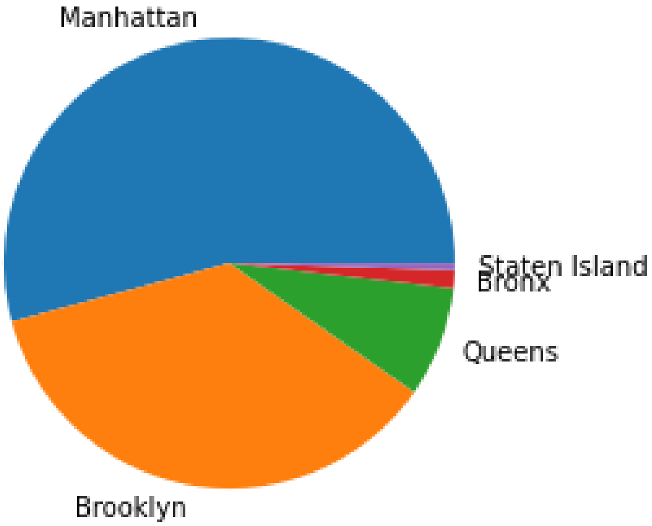


What can we learn about different hosts and areas.



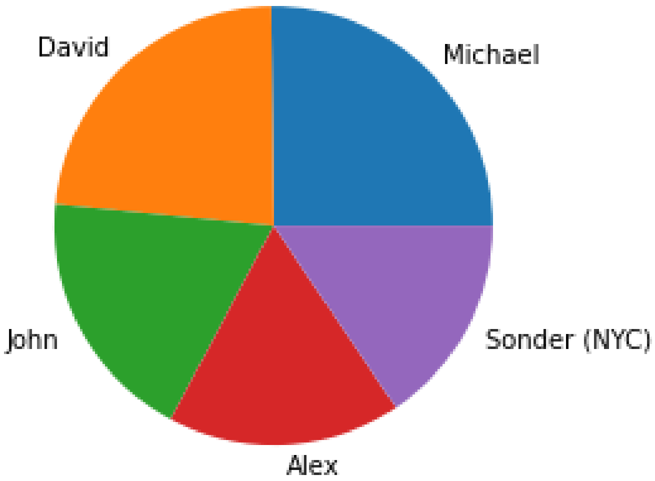
From the above analysis we get to know that all the top hosts are present in Williamsburg Neighbourhood and Manhattan Neighbourhood Group.

What can we learn from prediction.



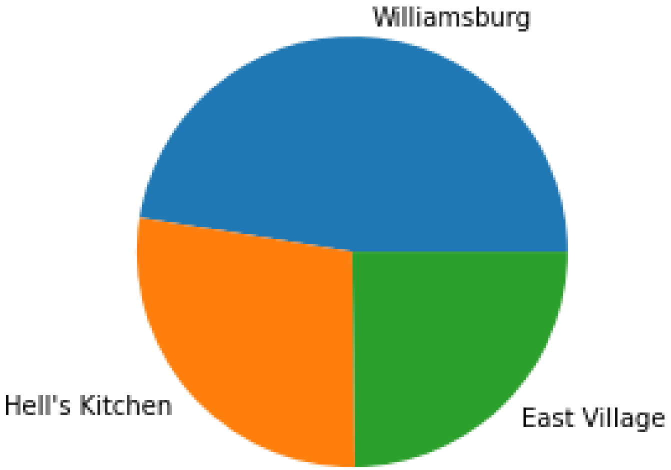
From the above price analysis we get to know that the Manhattan Neighbourhood group has highest price than other neighbourhood Groups.

Which hosts are the busiest and why.



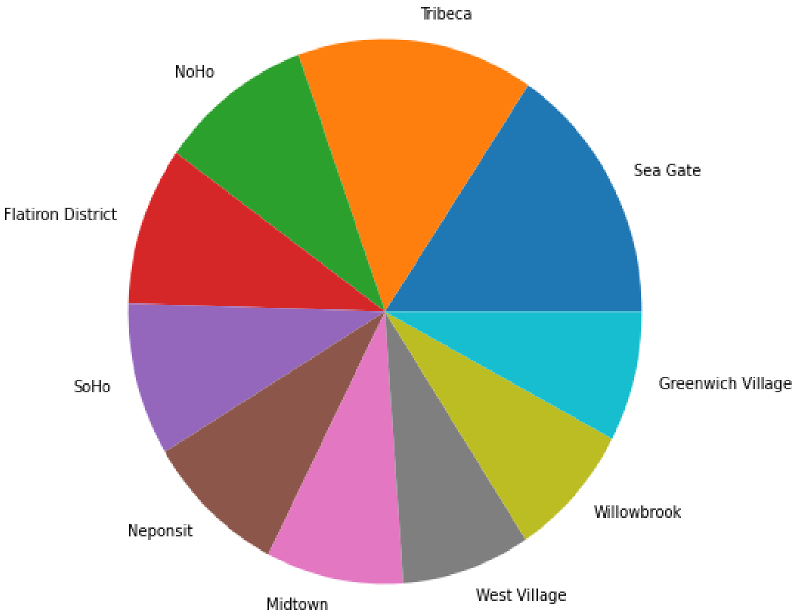
From the above graph we get to know that the Michael and david is busiest host than others becouse they are engage with more Neighbourhood Groups and Neighbourhoods and that's the reason behind it.

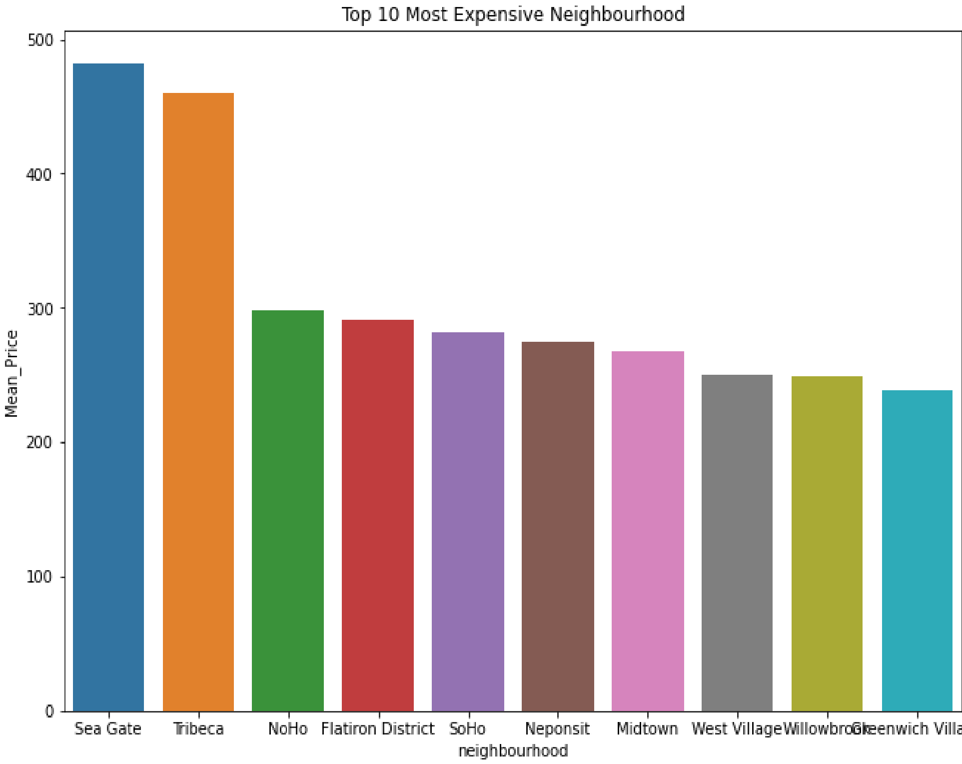
Is there any noticeable difference of traffic among different areas and what could be the reason for it ?



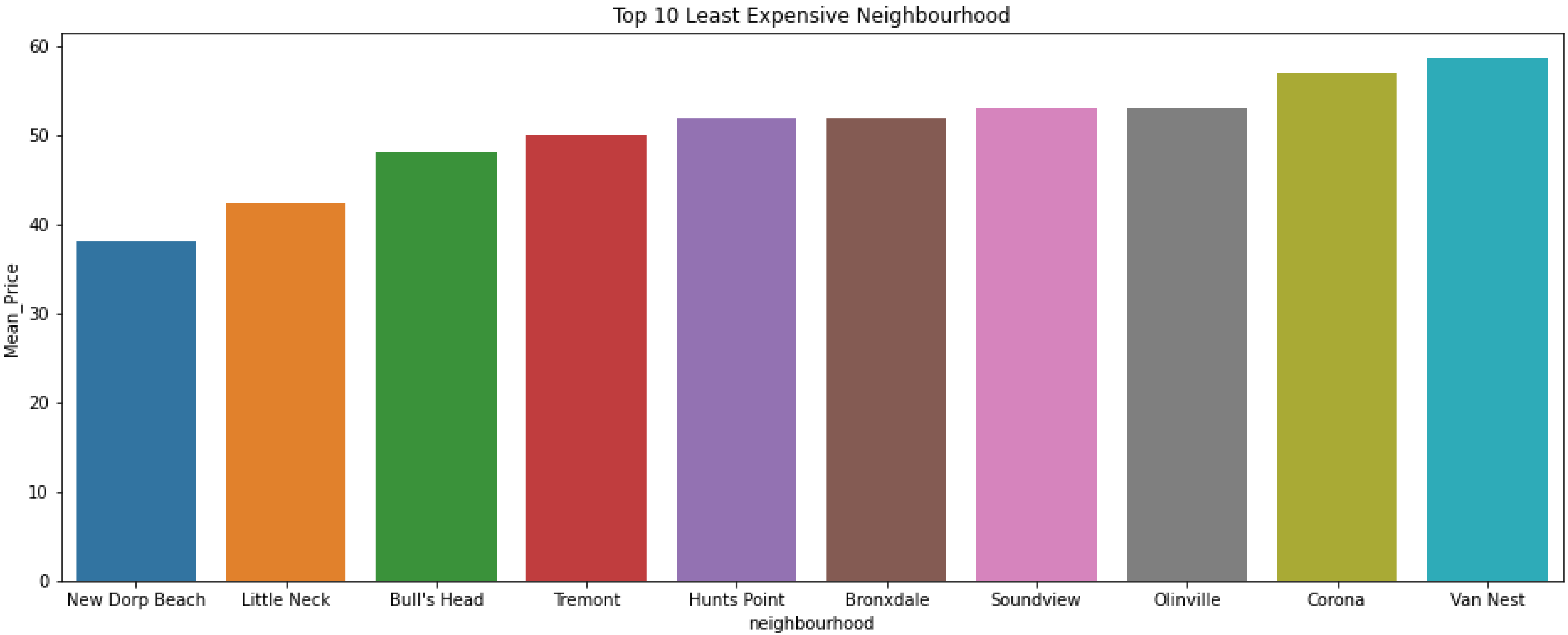
There is difference among the Neighbouhood reviews the reason behind it could be the price and quality provided by the host.

Top 10 Most Expensive Neighbourhood.





Top 10 Least Expensive Neighbourhood.



Conclusion

- **In this Exploratory Data Analysis we analyse the data of Airbnb with several key features such as price, neighbourhood, neighbourhood Group, Room type, number of reviews, etc.**

- We obtain price and neighbourhood relationship i.e., Manhattan is the most expensive airbnb region when we compare the other neighbourhood group. On the other hand the least expensive is region in Bronx.
- same analysis we did for the neighbourhood and through out that analysis we get to know that the most expensive neighbourhood is sea gate and the other hand the least expensive is new dorp Beach.
- Another analysis is conducted by using room type. The results show that the entire home/apt type is more preferable and the others are private room and shared room, respectively.
- In the host analysis we found that the Michael and David are the most busiest host.
- In the top host analysis we get to know that the top host are in Manhattan Neighbourhood Group and in williamsburg Neighbourhood.
- from the heatmap we get to know that the there is corelation between number of reviews and reviews per month
- Number of reviews are also investigated to find which neighborhoods take the most review according to the neighborhood group.

- Harshal Marathe